

# Delay-Optimal Biased User Association in Heterogeneous Networks

Fancheng Kong, Xinghua Sun, *Member, IEEE*, Victor C. M. Leung, *Fellow, IEEE*, and Hongbo Zhu

**Abstract**—In heterogeneous networks (HetNets), load balancing among different tiers can be effectively achieved by a biased user association scheme with which each user chooses to associate with one base station (BS) based on the biased received power. In contrast to previous studies where a BS always has packets to transmit, we assume in this paper that incoming packets intended for all the associated users form a queue in the BS. In order to find the delay limit of the network to support real-time service, we focus on the delay optimization problem by properly tuning the biasing factor of each tier. By adopting a thinned Poisson point process (PPP) model to characterize the locations of BSs in the busy state, an explicit expression of the average traffic intensity of each tier is obtained. On that basis, an optimization problem is formulated to minimize a lower bound of the network mean queuing delay. By showing that the optimization problem is convex, the optimal biasing factor of each tier can be obtained numerically. When the mean packet arrival rate of each user is small, a closed-form solution is derived. The simulation results demonstrate that the network queuing performance can be significantly improved by properly tuning the biasing factor. It is further shown that the network mean queuing delay might be improved at the cost of a deterioration of the network signal-to-interference ratio (SIR) coverage, which indicates a performance tradeoff between real-time and non-real-time traffic in HetNets.

**Index Terms**—heterogeneous network, biasing factor, average traffic intensity, network mean queuing delay

## I. INTRODUCTION

With widespread use of portable devices such as smart phones and tablets, cellular networks are facing an exponential growth of mobile data traffic [1]. Meanwhile, real-time applications such as video chat and online gaming become more and more popular, which imposes stringent delay requirements on the network. To deal with this ever-growing demand and high service requirement, micro base stations (BSs) such as pico and femto BSs are deployed to undertake the traffic pressure of macro BSs. The network architecture is thus evolving to more dense and irregular heterogeneous networks (HetNets) [2].

Among all the techniques used in HetNets, load balancing plays a key role to determine the network performance. For example, if the traditional maximum downlink received power association scheme is adopted, users would tend to connect to

macro BSs with a high transmission power. Macro BSs may thus easily become overloaded. To purposely push users to micro BSs, a simple and efficient approach called the biased association scheme was proposed in [3], with which each user assigns a biased value to the measured received power from BSs of each tier, and associates with the BS that has the largest mean biased received power. Ye *et al.* demonstrated in [4] that the user's long-term rate can be greatly improved by carefully tuning the biasing factor. With fixed locations of BSs and users, no tractable expression of a performance metric such as the signal-to-interference-plus-noise ratio (SINR) or rate coverage can be derived, and the optimal biasing factor could only be found empirically. Stochastic geometry was then adopted in [5]–[10] to characterize the spatial distributions of BSs and users, and to quantify the average performance metric of the network. For example, a Poisson point process (PPP) was adopted in [7], [8] to represent the irregular deployment of the BSs. The optimal biasing factor for BSs of each tier was obtained therein to maximize the rate coverage. With a similar PPP model, the biasing factor was optimized in [9], [10] by maximizing the logarithm of the mean user rate.

## A. Related Works and Motivations

Previous studies [4]–[10] assumed that the BSs always have packets to transmit, which presents a worst case for the SINR and rate coverage. In practice, the BS load could vary significantly over a day. In particular, the amount of user service requests can drop dramatically during non-peak traffic hours. The BSs are thus more likely to be idle during such periods, but still consume energy [11]. On the other hand, due to a high deployment density, one micro BS would cover a small area. The void cell problem [12], [13] then emerges, where some micro BSs don't have any associated users. Such BS thus solely acts as an interfering source. To improve energy efficiency and reduce the interference, the techniques to selectively switch off a fraction of BSs according to the traffic load have attracted extensive attention [14]–[18]. For example, the authors in [15] proposed a distributed on/off switching based algorithm in cellular networks to decide the minimum set of active BSs. By arguing that a cellular BS could operate in normal mode, sleep mode, or expansion mode, Guo *et al.* [16] proposed a scheme that determines which mode the BS should choose based on the load condition, such that the energy consumption is minimized. Dhillon *et al.* [17] adopted a thinned-PPP model by assigning an active probability to BSs of each tier and thus characterized the network signal-to-interference ratio (SIR) coverage. With the same model,

F. Kong, X. Sun, and H. Zhu are with the Jiangsu Key Laboratory of Wireless Communications, College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: kfeshimaidi@163.com; xinghua.sun.cn@ieee.org; zhb@njupt.edu.cn).

V. C. M. Leung is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: vleung@ece.ubc.ca).

Cao *et al.* [18] derived the optimal BS density of each tier to minimize the network energy consumption under a certain rate constraint.

The aforementioned studies [14]–[18] focused on the traffic load variance over a large time scale, i.e., peak and non-peak hours, and did not consider queuing in each BS. Most of them aimed to improve the network energy efficiency. One BS, nevertheless, can vary between busy and idle states over a small time scale due to the dynamic packet arrivals of its associated users, in which case the packet delay could be taken into account. In practice, with the proliferation of real-time multimedia applications, the packet delay is becoming an important quality-of-service (QoS) metric. For example, an end-to-end latency over 250 ms for real-time multimedia is generally considered to be unacceptable [19]. In the literature, there has been limited studies on the queuing analysis. The queuing performance of a single cell was evaluated in [20] and [21] in CDMA systems for the first time, based on which the packet blocking probability and the packet queuing delay were characterized. These studies focused on only one independent queue by assuming constant interference over the entire cell. In HetNets, nevertheless, BSs of various types are deployed with higher densities, and the queuing behavior of one BS is closely related with others, leading to the coupled-queue problem [17]. Specifically, whether a BS transmits or not will affect the interference level experienced by other BSs that occupy the same spectrum resources, and will consequently determine the service rates of these co-channel BSs. The service rate in turn affects the chance that one BS has packets to transmit or not.

The analysis of coupled queues is a long-standing open problem, and even solving a special case of two interacting queues is challenging [22]. Zhuang *et al.* [23] modeled multiple interacting queues as a continuous-time Markov chain (CTMC) with fixed BS locations in the network. By minimizing the average packet queuing delay, the optimal spectrum allocation pattern was obtained for each BS. Similarly, Cheng *et al.* [24] optimized the average queuing delay of network subjected to the BS power constraints. They formulated a Markov decision process (MDP) problem based on the instantaneous channel state information and queue state information, and proposed an adaptive user scheduling and power control policy. However, the state space of the Markov process may become huge as the network scales up, and the analysis would become intractable. Hence, this motivates us to deal with the coupled queue problem with the tool of stochastic geometry to account for the random BS deployment, and derive the mean performance metrics analytically such that some insights can be gained for system design.

## B. Our Approaches and Contributions

In this paper, we consider a  $K$ -tier HetNet where users and BSs of all tiers are randomly distributed, i.e., follow a PPP distribution. Similar to previous studies [5]–[10], it is assumed that each user adopts a biased association scheme to choose one BS with the maximum biased received power. In contrast to previous studies [4]–[10], [14]–[18], we consider that the packet requests from the users form a queue in their

associated BSs. The traffic intensity of one BS thus varies with the aggregate packet requests of all its associated users. To simplify the analysis, frequency partitioning across tiers is assumed in this paper. Although the queues of BSs from different tiers are not correlated, the queuing performance of one BS would affect the experienced interference of other co-channel BSs in the same tier. To decouple the queuing behavior of BSs in the same tier, we resort to the approximation of replacing each BS's individual traffic intensity with the average traffic intensity of its tier. The spatial distribution of BSs in the busy state can thus be approximately characterized by a thinned-PPP model [17]. The SIR coverage of each tier is then obtained, based on which an explicit expression of the average traffic intensity of each tier is further derived, and is shown to be an increasing function of the biasing factor of each tier.

In order to find the delay limit that the network can achieve, an optimization problem is formulated to minimize a lower bound of the network mean queuing delay by optimizing over the biasing factor of each tier based on the derived average traffic intensity. It is shown that the optimization problem is convex, and the optimal biasing factor can be numerically obtained. When the mean packet arrival rate of each user is small, an explicit expression of the optimal biasing factor of each tier is obtained. With equal bandwidth allocation across tiers, it is further shown that each user should associate with its nearest BS. A case study of a 2-tier HetNet shows that the optimal biasing factor is sensitive to the bandwidth allocation of each tier. To characterize the network capacity to support non-real-time services, the network SIR coverage is further derived. The contributions of this paper are summarized as follows.

- By assuming queuing in each BS, an explicit expression of the average traffic intensity of each tier is derived, which is shown to be an increasing function of the biasing factor of each tier.
- An optimization problem of a lower bound of the network mean queuing delay is formulated, and is shown to be convex with respect to the biasing factor of each tier. When the mean packet arrival rate of each user is small, an explicit solution is obtained.
- Simulation results of a 2-tier case demonstrate that the network mean queuing delay can be significantly reduced by properly tuning the biasing factor of each tier. In the meanwhile, a tradeoff is revealed between the network mean queuing delay and the network SIR coverage, which indicates that the service provider should strike a balance between the performance of real-time and non-real-time services.

The rest of this paper is organized as follows. The system model is presented in Section II. An optimization problem to minimize a lower bound of the network mean queuing delay is formulated and studied in Section III. A case study of a 2-tier HetNet is presented in Section IV. Conclusions and future works are given in Section V.

## II. SYSTEM MODEL

Consider a  $K$ -tier heterogeneous network where BSs in the  $k$ th tier form an independent PPP  $\Phi_k$  with an intensity of

$\lambda_k$ ,  $k \in \{1, \dots, K\}$ . Users, on the other hand, form another independent homogeneous PPP  $\Phi_u$  with an intensity of  $\lambda_u$  over the whole network. Frequency partitioning across tiers is assumed in this paper. In particular, BSs of the same tier share the spectrum with a bandwidth of  $W_k$ ,  $k \in \{1, \dots, K\}$ , and BSs of different tiers occupy non-overlapping frequency bands. Therefore, for each user in the downlink, the associated BS acts as a desired signal transmitter, and other BSs of the same tier are interfering sources. Consider a typical user located at the origin. Denote the distance between this typical user and a Tier- $k$  BS as  $x_k$ , and the transmission power of a Tier- $k$  BS as  $P_k$ . The received power  $P_R$  for a typical user from this BS can then be written as

$$P_R = P_k g_k x_k^{-\alpha}, \quad (1)$$

where  $g_k$  is a small-scale fading coefficient, which is assumed to follow an i.i.d exponential distribution of unit mean, i.e.,  $g_k \sim \exp\{1\}$ , and  $\alpha$  is the path-loss coefficient, which is assumed to be the same for all BSs in the network. Note that shadowing, i.e., log-normal fading, can be modeled by the randomness of the BSs' and users' locations [25].

In this paper, we consider a biased association scheme where users associate with one BS according to the maximum mean biased received power [4]–[10]. In particular, for a typical user located at the origin, it measures the mean received power from each tier's BSs, and chooses a Tier- $k$  BS if

$$P_k B_k x_{k,\min}^{-\alpha} \geq P_j B_j x_{j,\min}^{-\alpha} \quad \forall j \in \{1, \dots, K\}, \quad (2)$$

where  $B_j$  denotes the biasing factor of Tier  $j$  and  $x_{j,\min}$  is the distance between the user and the nearest Tier- $j$  BS.

For each user in the network, assume that its packet requests follow an independent Poisson process with a mean arrival rate  $\gamma$ , and the packet length is exponentially distributed with mean  $L$ . The incoming packets for all users form a queue in the associated BS, and the BS will transmit these packets in a first-in-first-serve (FIFS) fashion. To avoid users in poor channel conditions occupying the BS, we consider a fixed rate modulation and coding format. In particular, a BS will serve a user only when its instantaneous SIR exceeds a threshold  $\tau$ , and will drop its packet request otherwise. Note that due to a high BS deployment intensity, the background noise is dominated by the interference, and is therefore ignored in this paper. According to Shannon's formula, the service rate for each user that is associated to a Tier- $k$  BS can be obtained as

$$\mu_k = \frac{W_k}{L} \log_2(1 + \tau). \quad (3)$$

For a randomly selected Tier- $k$  BS,  $\text{BS}_{k,i}$ , its traffic intensity,  $\rho_{k,i}$ , can be obtained as

$$\rho_{k,i} = \frac{\gamma_{k,i}}{\mu_k}, \quad (4)$$

where  $\gamma_{k,i}$  is the mean aggregate packet arrival rate of all its associated users. Note that  $\rho_{k,i}$  can also be interpreted as the busy probability or the utilization of  $\text{BS}_{k,i}$  when  $\rho_{k,i} \leq 1$ . Due to a varied association region, each BS has a different mean aggregate packet arrival rate and thus has a different traffic intensity.

### III. QUEUING DELAY

In this section, we will formulate an optimization problem of the network mean queuing delay, which is an important performance metric of QoS. As the mean queuing delay of a BS increases with a higher busy probability, we will first characterize the average traffic intensity of each tier's BSs,  $\rho_k$ .

#### A. Average Traffic Intensity of Tier- $k$ BSs, $\rho_k$

For a randomly selected Tier- $k$  BS  $\text{BS}_{k,i}$ , since it delivers a packet only when the SIR exceeds a certain threshold  $\tau$ , its mean aggregate packet arrival rate can be obtained as

$$\gamma_{k,i} = \gamma N_{k,i} \Pr[\text{SIR}_{k,i} > \tau] \quad (5)$$

where  $N_{k,i}$  is the number of users that are associated to  $\text{BS}_{k,i}$  and  $\Pr[\text{SIR}_{k,i} > \tau]$  denotes the SIR coverage of  $\text{BS}_{k,i}$ , i.e., the probability that the SIR of a random user associated to  $\text{BS}_{k,i}$  is larger than the threshold  $\tau$ . By combining (3)–(5), the average traffic intensity of Tier- $k$  BSs can be obtained as

$$\begin{aligned} \rho_k &= E[\rho_{k,i}] = E\left[\frac{\gamma N_{k,i} \Pr[\text{SIR}_{k,i} > \tau]}{\mu_k}\right] \\ &= \frac{\gamma}{\mu_k} E[N_{k,i}] E[\Pr[\text{SIR}_{k,i} > \tau]] \\ &= \frac{\gamma L}{W_k \log_2(1 + \tau)} E[N_k] P[\text{SIR}_k > \tau], \end{aligned} \quad (6)$$

where  $E[N_k]$  denotes the average number of users associated with a Tier- $k$  BS and  $P[\text{SIR}_k > \tau]$  denotes the SIR coverage of all Tier- $k$  BSs, i.e., the probability that the SIR of a typical user associated with a Tier- $k$  BS exceeds the threshold  $\tau$ . As the average traffic intensity,  $\rho_k$ , is determined by the average number of associated users,  $E[N_k]$ , and the SIR coverage,  $P[\text{SIR}_k > \tau]$ , we will characterize these two components in the following.

According to [6], the average number of users associated with a Tier- $k$  BS,  $E[N_k]$ , has been obtained as

$$E[N_k] = \frac{\lambda_u A_k}{\lambda_k}, \quad (7)$$

where  $A_k$  denotes the probability for a typical user to be associated with a Tier- $k$  BS. Note that the association probability  $A_k$  has been derived in [6] as

$$A_k = \frac{\lambda_k (P_k B_k)^{2/\alpha}}{\sum_{j=1}^K \lambda_j (P_j B_j)^{2/\alpha}} = \frac{1}{\sum_{j=1}^K \tilde{\lambda}_j (\tilde{B}_j \tilde{P}_j)^{2/\alpha}}, \quad (8)$$

where  $\tilde{\lambda}_j = \lambda_j / \lambda_k$ ,  $\tilde{P}_j = P_j / P_k$ , and  $\tilde{B}_j = B_j / B_k$  denote the normalized intensity, the normalized transmission power, and the normalized biasing factor of Tier  $j$ , respectively, conditioned on Tier  $k$  being a serving tier.

Recall that BSs of Tier  $k$  form a PPP  $\Phi_k$  with an intensity of  $\lambda_k$ . Moreover, for a randomly selected  $\text{BS}_{k,i}$  where  $i \in \Phi_k$ , the traffic intensity  $\rho_{k,i}$  can be interpreted as its busy probability when  $\rho_{k,i} \leq 1$ . The set of Tier- $k$  BSs being busy, therefore, forms a thinned point process  $\Phi'_k \subseteq \Phi_k$  by including  $\text{BS}_{k,i} \in \Phi_k$  with the probability  $\rho_{k,i}$  [26]. Since the traffic intensity of

one BS is different from each other, the thinned point process  $\Phi'_k$  is non-homogeneous. To simplify the analysis, it can be approximately viewed as a homogeneous PPP with intensity

$$\lambda'_k = \rho_k \lambda_k. \quad (9)$$

It will be demonstrated in Section IV that the average traffic intensity achieves a good approximation. For a typical user that is associated with a Tier- $k$  BS, the interference all comes from busy BSs of the same tier. According to (1), the SIR of this typical user can then be written as

$$\text{SIR}_k = \frac{P_k g_{x_{k,0}} x_{k,0}^{-\alpha}}{\sum_{j \in \Phi'_k \setminus \text{BS}_{k,0}} P_k g_{k,j} x_{k,j}^{-\alpha}}, \quad (10)$$

where  $x_{k,0}$  and  $x_{k,j}$  denote the distance from the typical user to the associated BS  $\text{BS}_{k,0}$  and the  $j$ th interfering Tier- $k$  BS, respectively;  $g_{k,0}$  and  $g_{k,j}$  denote the small-scale fading coefficient of  $\text{BS}_{k,0}$  and the  $j$ th interfering Tier- $k$  BS, respectively. In (10),  $\text{BS}_{k,0}$  and  $\Phi'_k \setminus \text{BS}_{k,0}$  denote the associated Tier- $k$  BS of this typical user and the set of interfering Tier- $k$  BSs, respectively. Note that as frequency partitioning is assumed across tiers, there is no inter-tier interference, and the interfering sources consist of all the busy Tier- $k$  BSs besides the associated  $\text{BS}_{k,0}$ . Following a similar approach in [7], we have the following lemma that presents the SIR coverage of a Tier- $k$  BS.

**Lemma 1.** *The SIR coverage of a Tier- $k$  BS can be written as*

$$\text{P}[\text{SIR}_k > \tau] = \frac{1}{A_k \rho_k Z(\tau, \alpha, 1) + 1}, \quad (11)$$

where  $Z(\tau, \alpha, 1) = \tau^{\frac{2}{\alpha}} \int_{(1/\tau)^{\frac{2}{\alpha}}}^{\infty} \frac{du}{1+u^{\frac{2}{\alpha}}}$ , and the probability  $A_k$  for a typical user to be associated with a Tier- $k$  BS is given in (8).

According to Lemma 1, the outage probability of Tier- $k$ ,  $\text{P}[\text{SIR}_k \leq \tau]$ , can be written as  $\text{P}[\text{SIR}_k \leq \tau] = \frac{A_k \rho_k Z(\tau, \alpha, 1)}{A_k \rho_k Z(\tau, \alpha, 1) + 1}$ . If Tier- $k$  BSs are always busy, i.e.,  $\rho_k = 1$ , the outage probability  $\text{P}[\text{SIR}_k \leq \tau]$  reduces to [6, Eq. (17)].

By combining (6), (7), and (11), the average traffic intensity  $\rho_k$  of Tier- $k$  BSs can be derived as

$$\rho_k = \frac{-\lambda_k R_k + \left[ (\lambda_k R_k)^2 + 4\gamma \lambda_u \lambda_k A_k^2 R_k L Z \right]^{\frac{1}{2}}}{2A_k \lambda_k R_k Z} \quad (12)$$

where

$$R_k = W_k \log_2(1 + \tau), \quad (13)$$

and  $Z$  denotes  $Z(\tau, \alpha, 1)$  for simplicity. It is indicated in (12) that  $\rho_k$  is critically determined by the mean packet arrival rate of each user  $\gamma$  and the association probability  $A_k$ . It is clear that  $\rho_k$  increases as  $\gamma$  increases. On the other hand, the following lemma presents the monotonicity of the average traffic intensity  $\rho_k$  of Tier- $k$  BSs with respect to the association probability  $A_k$ .

**Lemma 2.** *The average traffic intensity  $\rho_k$  of Tier- $k$  BSs is an increasing function of its association probability,  $A_k$ .*

*Proof:* See Appendix B. ■

Intuitively, as the probability of a user being associated with a Tier- $k$  BS increases, more users from other tiers will be offloaded to BSs of Tier  $k$ , i.e.,  $E[N_k]$  becomes larger, which leads to an increment of the traffic intensity.

To this end, an explicit expression of  $\rho_k$  has been derived in (12). When the mean packet arrival rate of each user  $\gamma$  is small, the average traffic intensity  $\rho_k$  of Tier- $k$  BSs can be approximately written as

$$\begin{aligned} \rho_k &= \frac{-1 + \left[ 1 + 4\gamma \lambda_u A_k^2 (\lambda_k R_k)^{-1} L Z \right]^{\frac{1}{2}}}{2A_k Z} \\ &\stackrel{(a)}{\approx} \frac{-1 + 1 + 2\gamma \lambda_u A_k^2 (\lambda_k R_k)^{-1} L Z}{2A_k Z} \\ &= \frac{\gamma \lambda_u L A_k}{\lambda_k R_k}, \end{aligned} \quad (14)$$

where (a) follows from the fact that

$$\left[ 1 + \frac{4\gamma \lambda_u A_k^2 L Z}{\lambda_k R_k} \right]^{\frac{1}{2}} \approx 1 + \frac{2\gamma \lambda_u A_k^2 L Z}{\lambda_k R_k}. \quad (15)$$

## B. Queuing Delay Optimization

In this section, we will further minimize a lower bound of the network mean queuing delay based on the average traffic intensity by optimally tuning the biasing factors of all tiers. As each BS can be modeled as a M/D/1 queuing system, the mean queuing delay  $D_k$  of Tier  $k$  BSs can be obtained as

$$D_k = E \left[ \frac{L}{R_k (1 - \rho_{k,i})} \right]. \quad (16)$$

Since (16) is difficult to characterize, we resort to its lower bound using the convexity of  $1/(1 - \rho_{k,i})$ , i.e., we have

$$D_k \geq \bar{D}_k = \frac{L}{R_k (1 - E[\rho_{k,i}])} = \frac{L}{R_k (1 - \rho_k)}. \quad (17)$$

By combining (3), (12) and (17), the lower bound of the mean queuing delay of the whole network  $\bar{D}$  can then be written as (18), which is shown on the top of next page.

It can be observed from (18) that the lower bound of the mean queuing delay  $\bar{D}$  is critically determined by the association probability  $A_k$ . To minimize  $\bar{D}$ , we have the following optimization problem

$$\bar{D}^* = \underset{\{A_k\}_{\forall k \in \{1, \dots, K\}}}{\text{minimize}} \quad \bar{D}, \quad (19a)$$

$$\text{s.t.} \quad \sum_{k=1}^K A_k = 1, \quad (19b)$$

$$\rho_k < 1, \quad k \in \{1, \dots, K\}. \quad (19c)$$

Note that instead of directly optimizing over the biasing factor of each tier, we optimize over the association probabilities  $\{A_k\}_{\forall k}$  in (19) to obtain the optimal solution  $\{A_k^*\}_{\forall k}$ . The optimal normalized biasing factor of Tier  $k$  conditioned on

$$\bar{D} = \sum_{k=1}^K \frac{\lambda_k}{\sum_{j=1}^K \lambda_j} \cdot \bar{D}_k = \sum_{k=1}^K \frac{2A_k \lambda_k^2 LZ}{\sum_{j=1}^K \lambda_j \left( 2A_k \lambda_k Z R_k + \lambda_k R_k - \left[ (\lambda_k R_k)^2 + 4\gamma \lambda_u \lambda_k A_k^2 R_k LZ \right]^{\frac{1}{2}} \right)} \quad (18)$$

$$\begin{aligned} \frac{\partial \bar{D}}{\partial A_k} &= 2\lambda_K Z \frac{R_K \left( 1 - \sum_{j=1}^{K-1} A_j \right)^{-2} - R_K^2 \left( 1 - \sum_{j=1}^{K-1} A_j \right)^{-3} \left[ R_K^2 \left( 1 - \sum_{j=1}^{K-1} A_j \right)^{-2} + 4\gamma \lambda_u \lambda_K^{-1} R_K LZ \right]^{-\frac{1}{2}}}{\left\{ 2ZR_K + R_K \left( 1 - \sum_{j=1}^{K-1} A_j \right)^{-1} - \left[ R_K^2 \left( 1 - \sum_{j=1}^{K-1} A_j \right)^{-2} + 4\gamma \lambda_u \lambda_K^{-1} R_K LZ \right]^{\frac{1}{2}} \right\}^2} \\ &\quad - 2\lambda_k Z \frac{R_k A_k^{-2} - R_k^2 A_k^{-3} (R_k^2 A_k^{-2} + 4\gamma \lambda_u \lambda_k^{-1} R_k LZ)^{-\frac{1}{2}}}{\left[ 2ZR_k + R_k A_k^{-1} - (R_k^2 A_k^{-2} + 4\gamma \lambda_u \lambda_k^{-1} R_k LZ)^{\frac{1}{2}} \right]^2} = 0, \quad k \in \{1, \dots, K-1\}. \end{aligned} \quad (25)$$

Tier  $i$ ,  $\{\tilde{B}_k^*\}_{\forall k}$ , can then be readily obtained as

$$\tilde{B}_k^* = \frac{P_i(\lambda_i A_k^*)^{\frac{\alpha}{2}}}{P_k(\lambda_k A_i^*)^{\frac{\alpha}{2}}}, \quad k \in \{1, \dots, K\}, \quad (20)$$

according to (8). On the other hand, the constraint (19b) comes from the fact that each user should associate with one BS, and the constraint (19c) ensures that the lower bound of the network's mean queuing delay is bounded, which leads to the following lemma.

**Lemma 3.** *For the lower bound  $\bar{D}_k$ , when the mean packet arrival rate  $\gamma > \frac{(Z+1)\lambda_k R_k}{\lambda_u L}$ , it is bounded if the association probability*

$$0 < A_k < \frac{\lambda_k R_k}{\gamma \lambda_u L - \lambda_k R_k Z}; \quad (21)$$

*otherwise, it will become unbounded. When  $\gamma < \frac{(Z+1)\lambda_k R_k}{\lambda_u L}$ , it is always bounded.*

*Proof:* See Appendix C. ■

According to Lemma 3, constraint (19c) can be further written as

$$\begin{cases} 0 < A_k < \frac{\lambda_k R_k}{\gamma \lambda_u L - \lambda_k R_k Z}, & \gamma > \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \\ 0 < A_k < 1, & \gamma < \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \end{cases}, \quad (22)$$

where  $k \in \{1, \dots, K\}$ . First note that (22) does not have a feasible region if and only if

$$\gamma > \max_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \right\} \quad (23a)$$

and

$$\sum_{k=1}^K \frac{\lambda_k R_k}{\gamma \lambda_u L - \lambda_k R_k Z} < 1, \quad (23b)$$

according to (22). Intuitively, when the mean packet arrival rate of each user  $\gamma$  is too large, (22) can be written as  $0 < A_k < \frac{\lambda_k R_k}{\gamma \lambda_u L - \lambda_k R_k Z}$  for each Tier  $k$ ,  $k \in \{1, \dots, K\}$ ,

which leads to  $\sum_{k=1}^K A_k < 1$  according to (23b). In this case, the lower bound of the network mean queuing delay will always be unbounded. If (23) does not hold, the feasible region of the optimization problem (19) can be further written as

$$\begin{aligned} \mathbf{A} = & \left\{ (A_1, \dots, A_{K-1}), \left| 0 < A_j < \min \left\{ 1, \frac{\lambda_j R_j}{\gamma \lambda_u L - \lambda_j R_j Z} \right\}, \right. \right. \\ & j \in \{1, \dots, K-1\}; \max \left\{ 0, 1 - \frac{\lambda_K R_K}{\gamma \lambda_u L - \lambda_K R_K Z} \right\} < \sum_{j=1}^{K-1} A_j \\ & \left. < 1 \right\}, \end{aligned} \quad (24)$$

where  $A_K$  is eliminated according to the constraint (19b) without loss of generality. It is shown in Appendix D that the objective function (19a) is convex within the feasible region  $\mathbf{A}$ . Nevertheless, there may not exist a solution in  $\mathbf{A}$  through solving (25), which is shown at the top of this page, by setting the partial derivative of  $\bar{D}$  with respect to the association probability  $A_k$  to zero. The following lemma rules out this possibility and guarantees that the optimal association probabilities  $\{A_k^*\}_{\forall k}$  can always be obtained by finding the solution of (25) within  $\mathbf{A}$ .

**Lemma 4.** *(25) has a unique solution within the feasible region  $\mathbf{A}$ , which is the optimal association probabilities  $\{A_k^*\}_{\forall k}$ .*

*Proof:* See Appendix E. ■

So far we have demonstrated how to find the optimal association probability of each tier  $A_k^*$  by solving (25) numerically. Recall that it is indicated in Lemma 3 that when the mean packet arrival rate of each user  $\gamma < \min_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \right\}$ , we have the average traffic intensity  $\rho_k < 1$  for all tiers, and the lower bound of the mean queuing delay  $\bar{D}_k$  is always bounded for each tier. In this case, the average traffic intensity  $\rho_k$  is simply written as (14), and an explicit optimal association

**Algorithm 1** Procedure to optimize the association probability when the mean packet arrival rate of each user  $\gamma < \min_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \right\}$

- 1: **Input:**  $\lambda_k, W_k$  for each tier, and other system parameters  $\lambda_u, L, \gamma, \tau$ .
- 2: **Initialize:** a set of index  $C = \{1, \dots, K\}$  where optimal association probability of Tier  $k$  is not determined.
- 3: Calculate the solution set  $\{A_k^*\}_{\forall k \in C}$  by (26).
- 4: **for**  $\forall k \in C$ , construct a set  $S = \{m | A_m^* < 0, \forall m \in C\}$ .
- 5: **if**  $S = \emptyset$ , **return**  $\{A_k^*\}_{\forall k \in C}$ .
- 6: **else, for**  $\forall m \in S$ , let  $\lambda_m = 0$  and  $A_m^* = 0$ , delete  $m$  from  $C$ .
- 7: **end if**
- 8: **go to** Step 3.

probability  $A_k^*$  for each tier can be obtained, which is shown in the following lemma.

**Lemma 5.** When the mean packet arrival rate of each user  $\gamma < \min_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \right\}$ , the optimal association probability of Tier  $k$   $A_k^*$  to minimize the lower bound of the network mean queuing delay  $\bar{D}$  can be written as

$$A_k^* = \frac{\lambda_k}{\sum_{j=1}^K \lambda_j} + \frac{\lambda_k \log_2(1 + \tau) \sum_{j=1}^K \lambda_j (W_k - W_j)}{\gamma \lambda_u L \sum_{j=1}^K \lambda_j}. \quad (26)$$

The detailed derivation can be found in Appendix F.

Intuitively, if the bandwidth of Tier  $k$  is larger than that of Tier  $j$ , i.e.,  $W_k > W_j$ , the service rate of Tier  $k$  will be larger, indicating a better queuing performance. Therefore, the Tier- $k$  BSs will undertake more traffic from other tiers by having a larger association probability. With equal bandwidth allocation among all tiers, i.e.,  $W_i = W_j, i, k \in \{1, \dots, K\}$ , the optimal association probability of a Tier- $k$  BS can be further written as

$$A_k^* = \frac{\lambda_k}{\sum_{j=1}^K \lambda_j} \quad (27)$$

according to (26). The corresponding optimal normalized biasing factor  $\tilde{B}_k^*$  of Tier  $k$ , conditioned on Tier  $i$ , is thus given by

$$\tilde{B}_k^* = \frac{1}{\tilde{P}_k}, \quad (28)$$

where  $\tilde{P}_k$  is the normalized transmission power of Tier  $k$  conditioned on Tier  $i$ . It is indicated in (28) that in this case, each user chooses the nearest BS. The traffic load is thus evenly distributed among all BSs, which leads to similar queuing performance with the same service rate of each tier's BSs.

Note from (26) that if there exists one tier, say Tier  $m$ , such that  $\log_2(1 + \tau) \sum_{j=1}^K \lambda_j (R_m - R_j) < -\gamma \lambda_u L$ , and then we have

TABLE I  
SIMULATION PARAMETERS

Parameter	Value
User Density $\lambda_u$	$10^{-2} \text{ m}^{-2}$
Tier-1 BS Density $\lambda_1$	$10^{-4} \text{ m}^{-2}$
Tier-2 BS Density $\lambda_2$	$5 \times 10^{-4} \text{ m}^{-2}$
Tier-1 BS Transmission Power $P_1$	46 dBm
Tier-2 BS Transmission Power $P_2$	35 dBm
Path Loss Coefficient $\alpha$	4
Mean Packet Length $L$	0.1 Mb

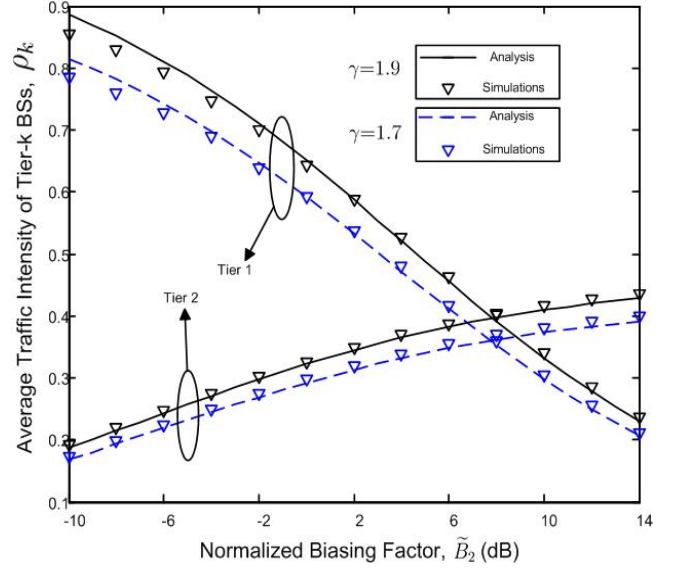


Fig. 1. Average traffic intensity of each tier  $\rho_k$  versus the normalized biasing factor  $\tilde{B}_2$  with various values of the mean packet arrival rate of each user  $\gamma$ .  $W_1 = 10\text{MHz}$ ,  $W_2 = 6\text{MHz}$ , and  $\tau = 1$ .

$A_m^* < 0$ . To minimize the lower bound of the network mean queuing delay, the association probability of Tier  $m$  should be close to zero. Intuitively, if the bandwidth of Tier  $m$  is much smaller than that of other tiers, then few users should associate with Tier- $m$  BSs due to the low service rate. In this case, the association probability  $A_m$  could then be simply set as  $A_m = 0$ , i.e., Tier- $m$  BSs are turned off. The procedure to obtain the optimal association probability when  $\gamma < \min_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \right\}$  is summarized in Algorithm 1.

#### IV. CASE STUDY

In this section we will demonstrate the analytical results in Section III by simulations of a 2-tier HetNet. The base stations and the users are drawn from PPPs with high intensities, and the background noise is ignored in the simulations. This setting, for example, can correspond to a dense heterogeneous network that consists of macro cellular BSs and micro Wi-Fi access points, each of which uses a non-overlapping frequency band. Each point of the simulation results is obtained by averaging all the BSs on a time scale of  $10^5$  s. The system parameters used in simulations are summarized in Table I.

Fig. 1 illustrates how the average traffic intensity of each tier, i.e.,  $\rho_1$  and  $\rho_2$ , varies with the normalized biasing factor

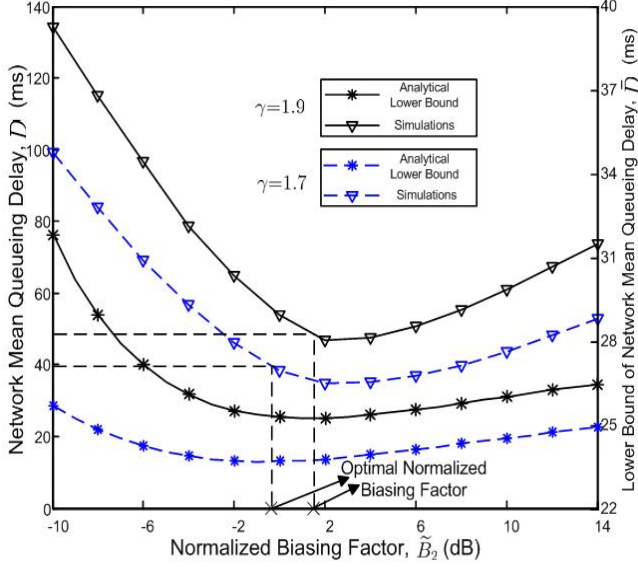


Fig. 2. Network mean queuing delay  $D$  and its lower bound  $\bar{D}$  versus the normalized biasing factor  $\tilde{B}_2$  with various values of the mean packet arrival rate of each user  $\gamma$ .  $W_1 = 10\text{MHz}$ ,  $W_2 = 6\text{MHz}$ , and  $\tau = 1$ .

$\tilde{B}_2$  with various values of the mean packet arrival rate of each user  $\gamma$ . It can be observed from Fig. 1 that the average traffic intensity of Tier 1,  $\rho_1$ , decreases as the normalized biasing factor  $\tilde{B}_2$  increases, while that of Tier 2,  $\rho_2$ , increases. Intuitively, since the association probability of a Tier-2 BS,  $A_2$ , increases as the normalized biasing factor  $\tilde{B}_2$  increases according to (8), more users that associate with Tier-1 BSs would be offloaded to Tier-2 BSs, which leads to an increment of  $\rho_2$  according to Lemma 2. Moreover, due to a larger deployment intensity of the Tier-2 BSs, the users that originally associate with only one Tier-1 BS can be offloaded to several neighboring Tier-2 BSs. Hence, the decline rate of  $\rho_1$  is larger than the increasing rate of  $\rho_2$ . It can be clearly seen from Fig. 1 that the simulation results match with the analysis well with a wide range of the normalized biasing factor, indicating that replacing each BS's traffic intensity by the average traffic intensity in (9) achieves a good approximation.

Fig. 2 further demonstrates how the network mean queuing delay  $D$ , as well as its lower bound  $\bar{D}$ , vary with the normalized biasing factor  $\tilde{B}_2$ . For the sake of comparison, the y-axis on the left hand side of Fig. 2 denotes the network mean queuing delay  $D$  while on the right hand side it denotes the lower bound  $\bar{D}$ . To obtain the network mean queuing delay in simulations, BSs that have an unbounded queuing delay are not taken account of. It can be observed from Fig. 2 that the trend of the network mean queuing delay  $D$  resembles that of its lower bound  $\bar{D}$ . Both  $D$  and  $\bar{D}$  are very sensitive to the normalized biasing factor  $\tilde{B}_2$ . If  $\tilde{B}_2$  is not carefully tuned, the delay performance could be greatly degraded. For example, when  $\gamma = 1.9$ , the network mean queuing delay  $D$  is as high as 135 ms with the normalized biasing factor  $\tilde{B}_2 = -10$  dB, which is not acceptable to many delay-sensitive applications. Moreover, due to a similar trend between the network mean queuing delay  $D$  and its lower bound  $\bar{D}$ , the optimal normalized biasing factor of  $\bar{D}$  is close

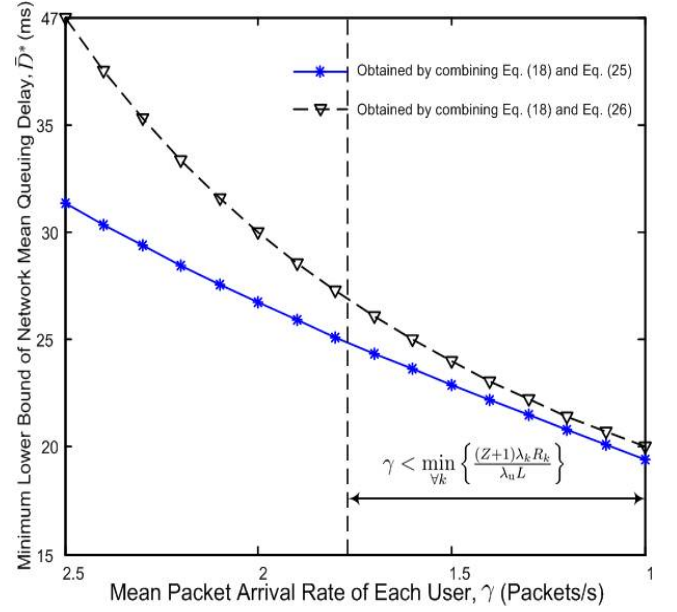


Fig. 3. Minimum lower bound of the network mean queuing delay  $\bar{D}^*$  versus the mean packet arrival rate of each user  $\gamma$ .  $W_1 = 10\text{MHz}$ ,  $W_2 = 6\text{MHz}$ , and  $\tau = 1$ .

to that of  $D$ . Therefore, by properly tuning the normalized biasing factor  $\tilde{B}_2$  according to (20) and (25), the mean queuing delay performance can be improved significantly. With the mean packet arrival rate of each user  $\gamma = 1.9$ , for instance, the optimal normalized biasing factor is obtained as  $\tilde{B}_2^* = 1.7$  dB, and the corresponding network mean queuing delay  $D$  can be reduced to be 48 ms.

Recall that it is indicated in Lemma 5 that when the mean packet arrival rate satisfies  $\gamma < \min_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \right\}$ , the minimum lower bound of the network mean queuing delay  $\bar{D}^*$  can be obtained by combining (18) and (26). Fig. 3 further compares the minimum lower bound of the network mean queuing delay  $\bar{D}^*$  obtained by combining (18) and (25) with that by combining (18) and (26), respectively. It can be observed from Fig. 3 that the gap between the two curves diminishes as the mean packet arrival rate of each user  $\gamma$  decreases. When  $\gamma < \min_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \right\}$ , the minimum lower bound of the network mean queuing delay  $\bar{D}^*$  obtained by combining (18) and (26) is quite close to that obtained by combining (18) and (25). When the mean packet arrival rate of each user  $\gamma$  is large, i.e.,  $\gamma \geq \min_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \right\}$ , there is a large gap between the curves in Fig. 3. Therefore, the optimal association probabilities  $\{A_k^*\}_{\forall k}$  should be instead obtained by numerically solving (25). As Lemma 4 guarantees, (25) has a unique solution within the feasible region  $\mathbf{A}$ , which is the optimal association probability  $\{A_k^*\}_{\forall k}$ .

Fig. 4 further illustrates how the optimal normalized biasing factor,  $\tilde{B}_2^*$ , and the corresponding minimum lower bound of the network mean queuing delay,  $\bar{D}^*$ , vary with the bandwidth ratio of Tier 2,  $W_2/W$ , with various values of the mean packet arrival rate of each user  $\gamma$ . Note that the total bandwidth  $W = W_1 + W_2$  is fixed here. It can be observed from Fig. 4(a) that for a given  $\gamma$ , the optimal normalized biasing factor  $\tilde{B}_2^*$



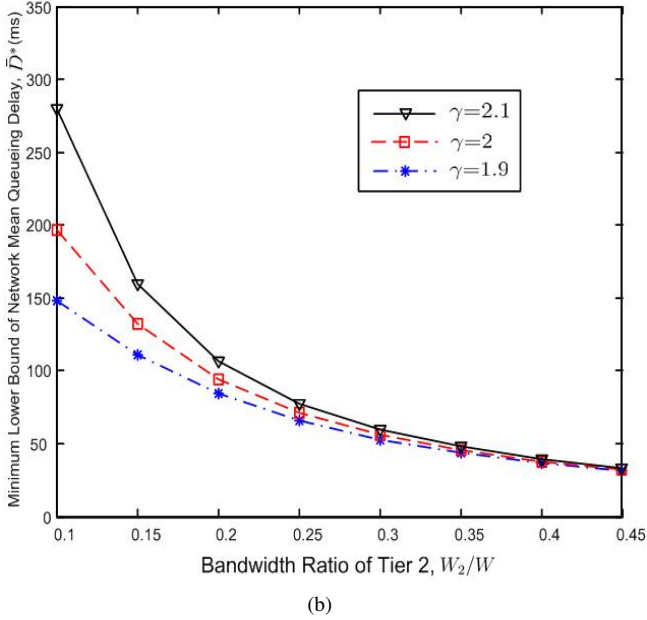
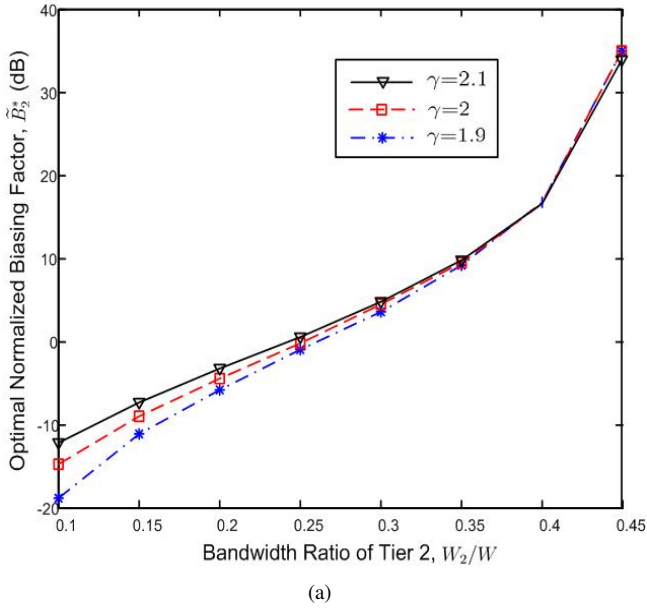


Fig. 4. Optimal normalized biasing factor  $\tilde{B}_2^*$  and the minimum lower bound of the network mean queuing delay  $\tilde{D}^*$  versus the bandwidth ratio of Tier 2  $W_2/W$  with various values of the mean packet arrival rate of each user  $\gamma$ .  $W=12\text{MHz}$  and  $\tau=1$ . (a) Optimal normalized biasing factor  $\tilde{B}_2^*$ . (b) Minimum lower bound of the network mean queuing delay  $\tilde{D}^*$ .

increases as  $W_2/W$  increases. Intuitively, as the bandwidth of Tier 2,  $W_2$ , increases, Tier-2 BSs can provide a higher service rate to the associated users. The optimal  $\tilde{B}_2^*$  should thus become larger so as to encourage more users to be associated with Tier-2 BSs. Moreover, it can be observed from Fig. 4(a) that as  $W_2/W$  increases, the optimal normalized biasing factor  $\tilde{B}_2^*$  becomes insensitive to the mean packet arrival rate of each user  $\gamma$ . The minimum lower bound of the network mean queuing delay  $\tilde{D}^*$ , on the other hand, decreases as  $W_2/W$  increases, as Fig. 4(b) demonstrates.

While minimizing the network mean queuing delay is desirable for real-time traffic, the SIR coverage is an important

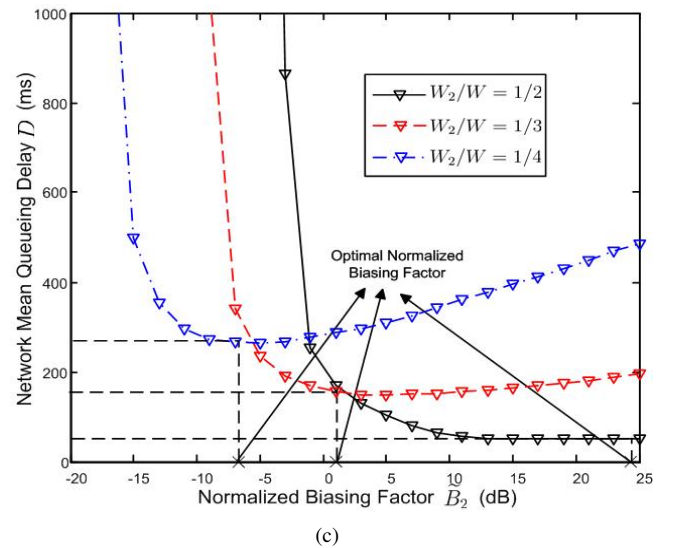
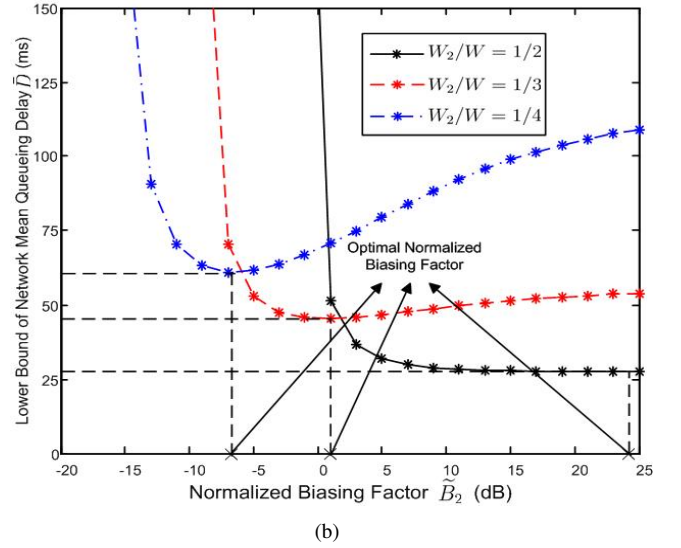
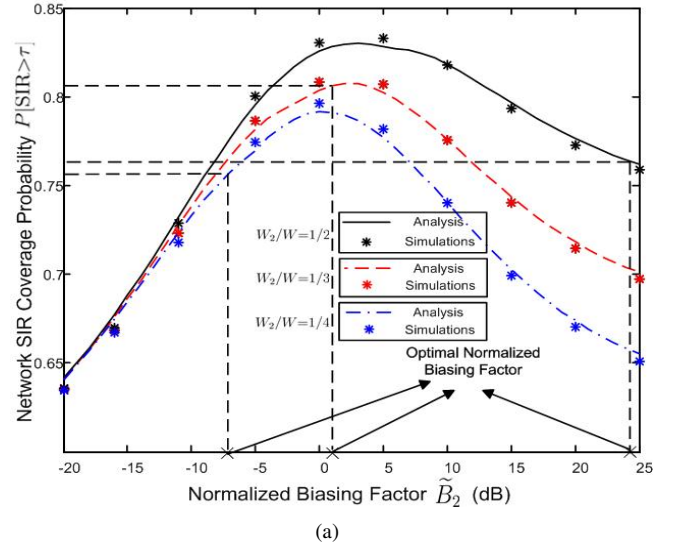


Fig. 5. The network SIR coverage and the network mean queuing delay performance with various bandwidth ratios of Tier 2  $W_2/W$ .  $\gamma=1.8$ ,  $W=12\text{MHz}$ , and  $\tau=1$ . (a) Network SIR coverage  $P[\text{SIR} > \tau]$ . (b) Network mean queuing delay  $\tilde{D}$ . (c) Lower bound of the network mean queuing delay  $\tilde{D}$ .



performance metric to support non-real-time traffic for service providers. According to (11), the network SIR coverage  $P[\text{SIR} > \tau]$  can be written as

$$\begin{aligned} P[\text{SIR} > \tau] &= \sum_{k=1}^K A_k \cdot P[\text{SIR}_k > \tau] \\ &= \sum_{k=1}^K \frac{A_k}{A_k \rho_k Z(\tau, \alpha, 1) + 1}. \end{aligned} \quad (29)$$

Fig. 5(a) demonstrates how the network SIR coverage  $P[\text{SIR} > \tau]$  varies with the normalized biasing factor  $\tilde{B}_2$  with various values of the bandwidth ratio  $W_2/W$ . It can be observed from Fig. 5(a) that there exists an optimal normalized biasing factor with which the network SIR coverage is maximized. Intuitively, when  $\tilde{B}_2$  is too large, a large fraction of users that are originally associated with Tier-1 BSs are offloaded to Tier-2 BSs. As these users are close to the interfering Tier-1 BSs and have long distances to their associated Tier-2 BSs, they have very poor channel conditions, which leads to a low SIR coverage of the network. Similarly, when  $\tilde{B}_2$  is too small, the network SIR coverage also deteriorates. In addition, it can be seen from Fig. 5(a) that the optimal normalized biasing to maximize  $P[\text{SIR} > \tau]$  is insensitive to the bandwidth allocation. In the meanwhile, the optimal normalized biasing factor  $\tilde{B}_2^*$  to minimize  $\bar{D}$  increases as  $W_2/W$  increases, as illustrated in Fig. 5(c) indicating a tradeoff between the network mean queuing delay and the network SIR coverage. For example, if  $W_2/W = 1/2$ , the optimal normalized biasing factor is obtained as  $\tilde{B}_2^* = 24\text{dB}$ , with which the network SIR coverage greatly deteriorates. In this case, the service providers should properly tune the biasing factor in HetNets such that a desired point on the tradeoff curve can be achieved to balance the performances of real-time traffic and non-real-time traffic.

As the SIR threshold  $\tau$  critically determines the network mean queuing delay and the network SIR coverage, Fig. 6 further demonstrates the impact of the SIR threshold  $\tau$  on these two performance metrics. It can be observed from Fig. 6 that for a given normalized biasing factor  $\tilde{B}_2$ , the network SIR coverage  $P[\text{SIR} > \tau]$  decreases as the SIR threshold  $\tau$  increases. In the meanwhile, both the network mean queuing delay  $\bar{D}$  and its lower bound  $\bar{\bar{D}}$  decrease as  $\tau$  increases. Intuitively, with a higher SIR threshold  $\tau$ , the mean aggregate packet arrival rate of each BS becomes lower while the service rate becomes higher, leading to a better queuing performance. In addition, it is illustrated in Fig. 6(a) that the optimal normalized biasing factor to maximize the network SIR coverage  $P[\text{SIR} > \tau]$  is insensitive to the SIR threshold  $\tau$ , while the optimal normalized biasing factor  $\tilde{B}_2^*$  to minimize  $\bar{D}$  increases as  $\tau$  decreases, as Fig. 6(c) demonstrates. Intuitively, although the service rates of both macro and micro BSs become lower with a smaller  $\tau$ , macro BSs are more likely to become overloaded as their deployment density is much lower than that of micro BSs. The optimal normalized biasing factor  $\tilde{B}_2^*$  should thus become larger to undertake the load pressure from macro BSs. By comparing Fig. 6(a) with Fig. 6(b) and Fig. 6(c), it can be found that with a smaller SIR threshold  $\tau$ , the

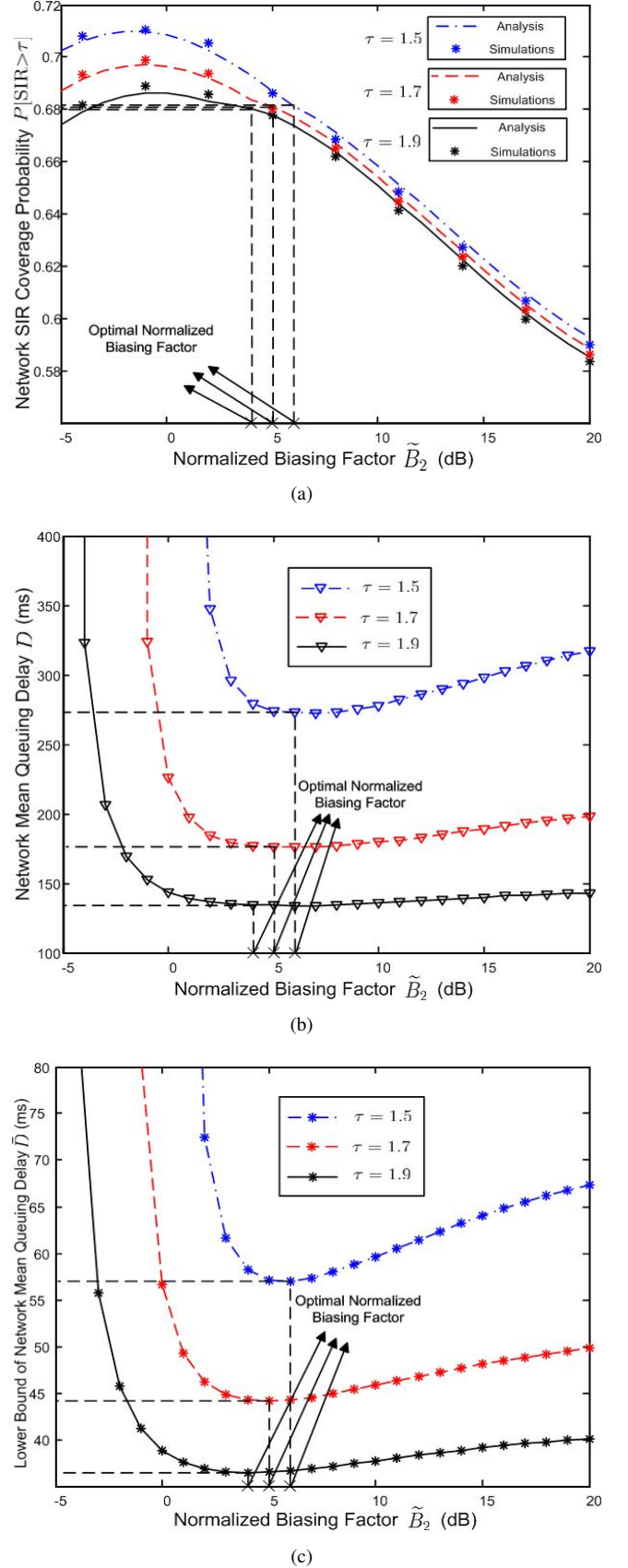


Fig. 6. The network SIR coverage and the network queuing delay performance with various values of the SIR threshold  $\tau$ .  $W_1 = 8\text{MHz}$ ,  $W_2 = 4\text{MHz}$ , and  $\gamma = 3.8$ , (a) Network SIR coverage  $P[\text{SIR} > \tau]$ . (b) Network mean queuing delay  $\bar{D}$ . (c) Lower bound of the network mean queuing delay  $\bar{\bar{D}}$ .

deterioration of the network mean queuing delay  $D$  becomes much more severe if the normalized biasing factor is optimally tuned to maximize the network SIR coverage  $P[\text{SIR} > \tau]$ , indicating a more significant tradeoff between the network SIR coverage and the network mean queuing delay.

## V. CONCLUSION AND FUTURE WORK

In this paper we have studied how to optimally tune the biasing factor of each tier in HetNets in order to minimize a lower bound of the network mean queuing delay. It is shown that the network queuing performance can be significantly improved when the biasing factor of each tier is optimally tuned. The characterization of the optimal biasing factor provides guidance for real-time service provisioning in HetNets. The case study of a 2-tier HetNet further illustrates that the network mean queuing delay and the network SIR coverage might not be optimized simultaneously by tuning the biasing factor, indicating a performance tradeoff between real-time and non-real-time services.

It is worth mentioning that it is assumed in this paper that one BS will serve a user with a constant rate if its SIR exceeds a threshold. In practice, nevertheless, the service rate could depend on the channel conditions. In this case, as the biasing factor of one tier decreases, the mean service rate of this tier increases as the users located at the edge of the cells are offloaded. The queuing performance of this tier can thus be improved due to a lower mean aggregate packet arrival rate and a higher mean service rate. Therefore, there would exist an optimal biasing factor for each tier such that the traffic load is balanced across tiers and the network mean queuing delay is minimized. On the other hand, if the biasing factor of one tier is too large, the SIR coverage of this tier degrades, which would drag down the network SIR coverage. Therefore, the network mean queuing delay may be optimized at the cost of the network SIR coverage. The tradeoff between the network SIR coverage and the network mean queuing delay in this case is an interesting issue that needs further study.

In addition, it is assumed that orthogonal spectrum resources are allocated to different tiers. In practice, nevertheless, universal frequency reuse may be adopted so that all the other BSs may act as interfering sources for one BS. Therefore, the average traffic intensities of different tiers would be correlated. The characterization of the queuing performance under such circumstances deserves much attention in future study.

### APPENDIX A PROOF OF LEMMA 2

*Proof:* According to (12), the first-order derivative of the average traffic intensity  $\rho_k$  with respect to  $A_k$  is given by

$$\frac{d\rho_k}{dA_k} = \frac{4\gamma L \lambda_u \lambda_k^2 R_k^2 A_k^2 Z^2 \Delta^{-\frac{1}{2}} - \lambda_k R_k Z \left( -\lambda_k R_k + \Delta^{\frac{1}{2}} \right)}{2(A_k \lambda_k R_k Z)^2}, \quad (30)$$

where  $\Delta = \lambda_k^2 R_k^2 + 4\gamma \lambda_u \lambda_k R_k A_k^2 L Z$ . The numerator on the right hand side of (30) can be further written as

$$4\gamma L \lambda_u \lambda_k^2 R_k^2 A_k^2 Z^2 \Delta^{-\frac{1}{2}} - \lambda_k R_k Z \left( -\lambda_k R_k + \Delta^{\frac{1}{2}} \right)$$

$$= \frac{\lambda_k^2 R_k^2 Z \left[ \left( \lambda_k^2 R_k^2 + 4\gamma \lambda_u \lambda_k R_k A_k^2 L Z \right)^{\frac{1}{2}} - \lambda_k R_k \right]}{\Delta^{\frac{1}{2}}} > 0. \quad (31)$$

By combining (30) and (31), we have  $\frac{d\rho_k}{dA_k} > 0$ , which indicates that  $\rho_k$  monotonically increases as  $A_k$  increases. ■

### APPENDIX B PROOF OF LEMMA 3

*Proof:* It has been shown in Lemma 2 that the average traffic intensity  $\rho_k$  monotonically increases as the association probability  $A_k$  increases. With  $A_k < 1$ , we then have

$$\begin{aligned} \rho_k &= \frac{-\lambda_k R_k + \left[ (\lambda_k R_k)^2 + 4\gamma \lambda_u \lambda_k R_k A_k^2 L Z \right]^{\frac{1}{2}}}{2A_k \lambda_k R_k Z} \\ &< \frac{-\lambda_k R_k + \left[ (\lambda_k R_k)^2 + 4\gamma \lambda_u \lambda_k R_k L Z \right]^{\frac{1}{2}}}{2\lambda_k R_k Z}. \end{aligned} \quad (32)$$

In the following, we divide the discussion into two parts:

1) If  $\frac{-\lambda_k R_k + \left[ (\lambda_k R_k)^2 + 4\gamma \lambda_u \lambda_k R_k L Z \right]^{\frac{1}{2}}}{2\lambda_k R_k Z} < 1$ , i.e.,  $\gamma < \frac{(Z+1)\lambda_k R_k}{\lambda_u L}$ , we have

$$\rho_k < 1 \quad (33)$$

according to (32). In this case,  $\bar{D}_k$  will always be bounded if  $\gamma < \frac{(Z+1)\lambda_k R_k}{\lambda_u L}$ .

2) If  $\gamma > \frac{(Z+1)\lambda_k R_k}{\lambda_u L}$ ,  $\bar{D}_k$  will be bounded if and only if

$$\frac{-\lambda_k R_k + \left[ (\lambda_k R_k)^2 + 4\gamma \lambda_u \lambda_k A_k^2 R_k L Z \right]^{\frac{1}{2}}}{2A_k \lambda_k R_k Z} < 1. \quad (34)$$

Accordingly, we have

$$A_k < \frac{\lambda_k R_k}{\gamma \lambda_u L - \lambda_k R_k Z}. \quad (35)$$

■

### APPENDIX C PROOF OF CONVEXITY OF (19)

*Proof:* According to (17), the second-order derivative of  $\bar{D}_k$  with respect to  $A_k$  can be written as

$$\frac{d^2 \bar{D}_k}{dA_k^2} = \frac{2L}{R_k (1 - \rho_k)^3} \cdot \left( \frac{d\rho_k}{dA_k} \right)^2 + \frac{L}{R_k (1 - \rho_k)^2} \cdot \frac{d^2 \rho_k}{dA_k^2}. \quad (36)$$

Substituting (30) into (36) yields

$$\begin{aligned} \frac{d^2 \bar{D}_k}{dA_k^2} &> \frac{L}{R_k (1 - \rho_k)^2} \cdot \left[ 2 \left( \frac{d\rho_k}{dA_k} \right)^2 + \frac{d^2 \rho_k}{dA_k^2} \right] \\ &= \frac{L}{R_k (1 - \rho_k)^2 A_k^4 Z^2 \Delta} \cdot \left( 4\gamma \lambda_u L \lambda_k^2 R_k^2 Z^2 A_k^3 + 2\Delta \right) \end{aligned}$$

$$\begin{aligned}
& + 2\lambda_k R_k A_k Z \Delta^{\frac{1}{2}} + \lambda_k^2 R_k^2 - 2A_k Z \Delta - \lambda_k R_k \Delta^{\frac{1}{2}} \Big) \\
& > \frac{L}{R_k (1 - \rho_k)^2 A_k^4 Z^2 \Delta} \cdot \left[ 4\gamma \lambda_u L \lambda_k^2 R_k^2 Z^2 A_k^3 \right. \\
& \quad \left. + \lambda_k R_k \left( 2A_k Z \Delta^{\frac{1}{2}} + \lambda_k R_k - \Delta^{\frac{1}{2}} \right) \right], \quad (37)
\end{aligned}$$

where  $\Delta = \lambda_k^2 R_k^2 + 4\gamma \lambda_u \lambda_k R_k A_k^2 L Z$ . Since  $\Delta^{\frac{1}{2}} > \lambda_k R_k$ , we further have

$$\begin{aligned}
\frac{d^2 \bar{D}_k}{dA_k^2} & > \frac{L}{R_k (1 - \rho_k)^2 A_k^4 Z^2 \Delta} \cdot \left[ 4\gamma \lambda_u L \lambda_k^2 R_k^2 Z^2 A_k^3 \right. \\
& \quad \left. + \lambda_k R_k \left( 2A_k Z \Delta^{\frac{1}{2}} + \lambda_k R_k - \Delta^{\frac{1}{2}} \right) \right] \\
& > \frac{L}{R_k (1 - \rho_k)^2 A_k^4 Z^2 \Delta} \cdot \left[ 4\gamma \lambda_u L \lambda_k^2 R_k^2 Z^2 A_k^3 \right. \\
& \quad \left. + \lambda_k R_k \left( 2\lambda_k R_k A_k Z + \lambda_k R_k - \Delta^{\frac{1}{2}} \right) \right] \\
& \stackrel{(a)}{>} \frac{4\gamma \lambda_u L^2 \lambda_k^2 R_k}{(1 - \rho_k)^2 A_k \Delta} > 0, \quad (38)
\end{aligned}$$

where (a) follows from the fact that  $\rho_k < 1$ . As the constraints (19b) and (19c) are linear, it can be concluded from (38) that the optimization problem is convex with respect to  $A_k$ . ■

#### APPENDIX D PROOF OF LEMMA 4

*Proof:* We divide the proof into two parts.

- 1) If  $\gamma > \max_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \right\}$ , then the mean queuing delay  $\bar{D}_k$  of all tiers go to infinity as  $A_k$  approaches to 1. Therefore, according to (18), the lower bound of the network mean queuing delay,  $\bar{D}$ , goes to infinity at the boundary of  $\mathbf{A}$ . As  $\bar{D}$  is convex within the region  $\mathbf{A}$ , (25) always has a unique solution of the optimal association probabilities  $\{A_k^*\}_{\forall k}$ .
- 2) If  $\gamma < \max_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \right\}$ , then there exists at least one tier such that the lower bound of its mean queuing delay is always bounded. Without loss of generality, denote this tier as Tier  $K$ . For Tier  $K$ , we have  $\frac{\lambda_K R_K}{\gamma \lambda_u L - \lambda_K R_K Z} > 1$ , and the feasible region  $\mathbf{A}$  is then written as

$$\begin{aligned}
\mathbf{A} = & \left\{ (A_1, \dots, A_{K-1}), \left| 0 < A_k < \min \left\{ 1, \frac{\lambda_k R_k}{\gamma \lambda_u L - \lambda_k R_k Z} \right\}, \right. \right. \\
& \left. \left. k \in \{1, \dots, K-1\}; 0 < \sum_{k=1}^{K-1} A_k < 1 \right\}. \quad (39)
\end{aligned}$$

For each  $k \in \{1, \dots, K-1\}$ , we have

$$\lim_{A_k \rightarrow 0} \frac{\partial \bar{D}}{\partial A_k} = 2\lambda_K Z.$$

$$\frac{R_K A_K^{-2} \left[ 1 - (1 + 4\gamma \lambda_u \lambda_K^{-1} A_K^2 R_K^{-1} L Z)^{-\frac{1}{2}} \right]}{\left[ 2Z R_K + R_K A_K^{-1} - (R_K^2 A_K^{-2} + 4\gamma \lambda_u \lambda_K^{-1} R_K L Z)^{\frac{1}{2}} \right]^2} < 0 \quad (40)$$

according to (25).

Following a similar approach, if  $\frac{\lambda_k R_k}{\gamma \lambda_u L - \lambda_k R_k Z} > 1$ , we have

$$\lim_{A_k \rightarrow 1} \frac{\partial \bar{D}}{\partial A_k} > 0. \quad (41)$$

Otherwise, if  $\frac{\lambda_k R_k}{\gamma \lambda_u L - \lambda_k R_k Z} < 1$ , the lower bound  $\bar{D}_k$  goes to infinity as  $A_k$  approaches  $\frac{\lambda_k R_k}{\gamma \lambda_u L - \lambda_k R_k Z}$ , and thus we have

$$\lim_{A_k \rightarrow \frac{\lambda_k R_k}{\gamma \lambda_u L - \lambda_k R_k Z}} \frac{\partial \bar{D}}{\partial A_k} > 0. \quad (42)$$

By combining (40)-(42), it can be concluded that (25) always has only one solution within the region  $0 < A_k < \min\{1, \frac{\lambda_k R_k}{\gamma \lambda_u L - \lambda_k R_k Z}\}$ ,  $k \in \{1, \dots, K-1\}$ .

Furthermore, if  $\sum_{k=1}^{K-1} A_k > 1$ , i.e.,  $A_K < 0$ , we always have  $\frac{\partial \bar{D}}{\partial A_k} > 0$ ,  $k \in \{1, \dots, K-1\}$  by substituting  $A_K < 0$  into (25). This indicates that the solution is not in the region where  $\sum_{k=1}^{K-1} A_k > 1$ . Therefore, (25) has a unique solution in region  $\mathbf{A}$  when  $\gamma < \max_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \right\}$ . ■

#### APPENDIX E DERIVATION OF (26)

By combining (14), (18), and (19b), when the mean packet arrival rate of each user satisfies  $\gamma < \min_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \right\}$ , the lower bound of the network mean queuing delay can be written as

$$\begin{aligned}
\bar{D} = & \frac{1}{\sum_{j=1}^K \lambda_j} \sum_{k=1}^K \frac{\lambda_k^2 L}{\lambda_k R_k - \gamma \lambda_u L A_k} = \frac{1}{\sum_{j=1}^K \lambda_j} \left[ \right. \\
& \left. \sum_{k=1}^{K-1} \frac{\lambda_k^2 L}{\lambda_k R_k - \gamma \lambda_u L A_k} + \frac{\lambda_K^2 L}{\lambda_K R_K - \gamma \lambda_u L (1 - \sum_{j=1}^{K-1} A_j)} \right] \quad (43)
\end{aligned}$$

where  $R_k$  is given by (13). By setting the partial derivative of  $\bar{D}$  with respect to  $A_k$  to zero, we have

$$\begin{aligned}
\frac{\partial \bar{D}}{\partial A_k} = & \frac{\lambda_k^2}{\sum_{j=1}^K \lambda_j} \cdot \frac{\lambda_u \gamma}{(\lambda_k \frac{R_k}{L} - \lambda_u \gamma A_k)^2} - \frac{\lambda_K^2}{\sum_{j=1}^K \lambda_j} \\
& \frac{\lambda_u \gamma}{\left[ \lambda_k \frac{R_k}{L} - \lambda_u \gamma \left( 1 - \sum_{j=1}^{K-1} A_j \right) \right]^2} = 0, \quad \forall k \in \{1, \dots, K-1\}. \quad (44)
\end{aligned}$$

By combining (19b) and (44), (26) can be obtained.

## REFERENCES

- [1] J. Andrews, S. Buzzi, C. Wan, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, June 2014.
- [2] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in hetnets: old myths and open problems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 18–25, Apr. 2014.
- [3] H. M. Carlos, B. Mehdli, and L. Matti, "Statistical analysis of self-organizing networks with biased cell association and interference avoidance," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 1950–1961, Feb. 2013.
- [4] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, June 2013.
- [5] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Apr. 2012.
- [6] H. Jo, Y. Sang, P. Xia, and J. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.
- [7] S. Singh, H. Dhillon, and J. Andrews, "Offloading in heterogeneous networks: modeling, analysis, and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, May 2013.
- [8] S. Singh, and J. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 888–901, Feb. 2014.
- [9] Y. Lin, W. Bao, W. Yu, and B. Liang, "Optimizing user association and spectrum allocation in HetNets: a utility perspective," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1025–1039, Mar. 2015.
- [10] F. Kong, X. Sun, and H. Zhu, "Optimal biased association scheme with heterogeneous user distribution in HetNets," *Wireless Personal Commun.*, vol. 90, no. 2, pp. 575–594, Sept. 2016.
- [11] J. Wu, Y. Zhang, M. Zukerman, and E. K. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: a survey," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 2, pp. 803–826, 2015.
- [12] C. H. Liu and L. C. Wang, "Random cell association and void probability in Poisson-distributed cellular networks," in *Proc. IEEE ICC*, June 2015, pp. 2816–2821.
- [13] C. T. Peng, L. C. Wang and C. H. Liu, "Optimal base station deployment for small cell networks with energy-efficient power control," in *Proc. IEEE ICC*, June 2015, pp. 1863–1868.
- [14] H. Jung, H. Roh, J. Lee, "Energy and traffic aware dynamic topology management for wireless cellular networks," in *Proc. IEEE ICCS*, 2012, pp. 205–209.
- [15] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2126–2136, 2013.
- [16] W. Guo, T. Farrell, "Dynamic cell expansion: traffic aware low energy cellular network," in *Proc. IEEE VTC*, Sept. 2012, pp. 1–5.
- [17] H. Dhillon, R. Ganti, and J. Andrews, "Load-aware modeling and analysis of heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1666–1677, Apr. 2013.
- [18] D. Cao, S. Zhou, and Z. Niu, "Optimal combination of base station densities for energy-efficient two-tier heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4305–4362, Apr. 2013.
- [19] W. Tu, and W. Jia, "Adaptive playback buffer for wireless streaming media," in *Proc. IEEE ICON*, 2004, pp. 191–195.
- [20] T. Bonald and A. Proutière, "Wireless downlink data channels: user performance and cell dimensioning," in *Proc. ACM MOBICOM*, 2003, pp. 339–352.
- [21] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc. IEEE INFOCOM*, 2003, pp. 321–331.
- [22] I. J. B. F. Adan, J. Wessels, and W. H. M. Zijm, "A compensation approach for two-dimensional Markov processes," *Adv. Appl. Probab.*, vol. 25, no. 4, pp. 783–817, Dec. 1993.
- [23] B. Zhuang, D. Guo, and M. L. Honig, "Traffic-driven spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2027–2038, Oct. 2015.
- [24] A. Cheng, J. Li, Y. Yu, and H. Jin, "Delay-sensitive user scheduling and power control in heterogeneous networks," *IET Networks*, vol. 4, no. 3, pp. 175–184, 2015.
- [25] B. Błaszczyszyn, M. K. Karray, and H. P. Keeler, "Using Poisson processes to model lattice cellular networks," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 773–781.
- [26] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications.*, John Wiley & Sons, 1996.